# SAM TOOLS TIME

Benni Goetz
CCBB Bioinformatics Consultant
Feb 19, 2014

Q: What does Samtools do?

A: Lets you view and manipulate SAM/BAM files.

Q: Awesome.

Q: What are SAM/BAM files?

# SAM Files

```
@SQ     SN:NC_012967    LN:4629812
@PG     ID:bwa  PN:bwa  VN:0.7.4-r385
SRR030257.1     99      NC_012967       950180  60      36M     =       950295  151     TTACACTCCTGTTAATCCATACAGCAACAGTATTGG
AAA;A;AA?A?AAAAA?;?A?1A;;????566)=*1    XT:A:U  NM:i:1  SM:i:37 AM:i:25 X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:32C3
```

- The @ lines are headers. That's metadata you don't normally need to deal with.
- The next two lines are actually a single line in the SAM file, representing an alignment of paired-end reads to *E.coli*.

| QNAME | FLAG | RNAME | POS | MAPQ | CIGAR | RNEXT | PNEXT | TLEN | SEQ |
|---|---|---|---|---|---|---|---|---|---|
| SRR030257.1 | 99 | NC_012967 | 950180 | 60 | 36M | = | 950295 | 151 | TTACACTCCTGTTAATCCATACAGCAACAGTATTGG |

- QNAME: Name of the read.
- FLAG: Metadata (is the read part of paired-end data? did it map?)
- RNAME: Name of the contig that it mapped to.
- POS: Position of RNAME that the read mapped to (from the left)
- MAPQ: Phred-scaled mapping quality.
- CIGAR: Summary of matches, insertions, deletions. (5M2I31M would indicate a 5bp (mis-)match, a 2bp insertion, followed by a 31bp (mis-)match.)
- RNEXT: Name of the contig the mate pair aligns to.
- PNEXT: Position of RNEXT that mate pair aligns to.
- TLEN: Only applies to paired-end reads. Estimated total length of the read, calculated from the left-most base to the right-most base from both reads.
- SEQ: The original sequence itself!

**QUAL** **BUNCH O' METADATA**

```
AAA;A;AA?A?AAAAA?;?A?1A;;????566)=*1    XT:A:U  NM:i:1  SM:i:37 AM:i:25 X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:32C3
```

- QUAL: Quality score for each base, taken from the original FASTQ file.

- The rest of the metadata is optional, and different mappers may add different metadata. There are a few conventions, however.

- NM: Number of base pair changes required to match read with reference. NM:i:1 indicates that one base was changed to match the reference.

- MD: Summary of mismatch positions. MD:Z:32C3 means that 32bp matched the reference, then the reference contains a C where the read doesn't, and then the remaining 3bp match.

# About FLAG

| Flag | Chr | Description |
|---|---|---|
| 0x0001 | p | the read is paired in sequencing |
| 0x0002 | P | the read is mapped in a proper pair |
| 0x0004 | u | the query sequence itself is unmapped |
| 0x0008 | U | the mate is unmapped |
| 0x0010 | r | strand of the query (1 for reverse) |
| 0x0020 | R | strand of the mate |
| 0x0040 | 1 | the read is the first read in a pair |
| 0x0080 | 2 | the read is the second read in a pair |
| 0x0100 | s | the alignment is not primary |
| 0x0200 | f | the read fails platform/vendor quality checks |
| 0x0400 | d | the read is either a PCR or an optical duplicate |

# What?

| Flag | Chr | Description |
|---|---|---|
| 0x0001 | p | the read is paired in sequencing |
| 0x0002 | P | the read is mapped in a proper pair |
| 0x0004 | u | the query sequence itself is unmapped |
| 0x0008 | U | the mate is unmapped |
| 0x0010 | r | strand of the query (1 for reverse) |
| 0x0020 | R | strand of the mate |
| 0x0040 | 1 | the read is the first read in a pair |
| 0x0080 | 2 | the read is the second read in a pair |
| 0x0100 | s | the alignment is not primary |
| 0x0200 | f | the read fails platform/vendor quality checks |
| 0x0400 | d | the read is either a PCR or an optical duplicate |

- FLAG is a sequence of bits that are on/off depending on how the read mapped.
- 00000000101 would mean that the read has a pair, and that the read didn't map to the reference. But this is binary representation for the number 3. So the FLAG field would show a 3.
- To interpret the 99 in our example line, write 99 in binary and check the sequence of 0's and 1's. Or use http://picard.sourceforge.net/explain-flags.html to decode the FLAG. (Google "Explain SAM Flags".)

# BAM Files

- Binary representation of SAM files.
- Much smaller, faster for software to work with. Samtools requires BAM files for many of its operations.
- Easy for robots to read, impossible for humans.
  (No Voight-Kampff machine in your lab? Casually ask your lab mate to read a BAM file. If she/he/it can, it's a replicant.)

- `samtools view -bS file.sam > file.bam`
- `samtools view file.bam`
- `samtools flagstat`

# Extract (un)mapped reads

- Extract unmapped reads:
  samtools view -f 4 file.bam > unmapped.sam

- Extract mapped reads:
  samtools view -F 4 file.bam > mapped.sam

- Extract unmapped reads using bioawk:
  bioawk -c sam 'and($flag,4)' file.sam

# Other Common Uses

- Merging BAM files: use samtools merge

- Variant-calling: use “samtools sort” and “samtools index” on your BAM files, then use “samtools mpileup” to generate a bcf file.

# Go Forth

- SAM files encode all the relevant data for reads and their alignments in a tab-delimited text file. Can store a binary file to save space, then generate FASTA, FASTQ, whatever files as needed.

- Lots of information online: check the BioITeam wiki, google “samtools tutorial”.